Data analysis

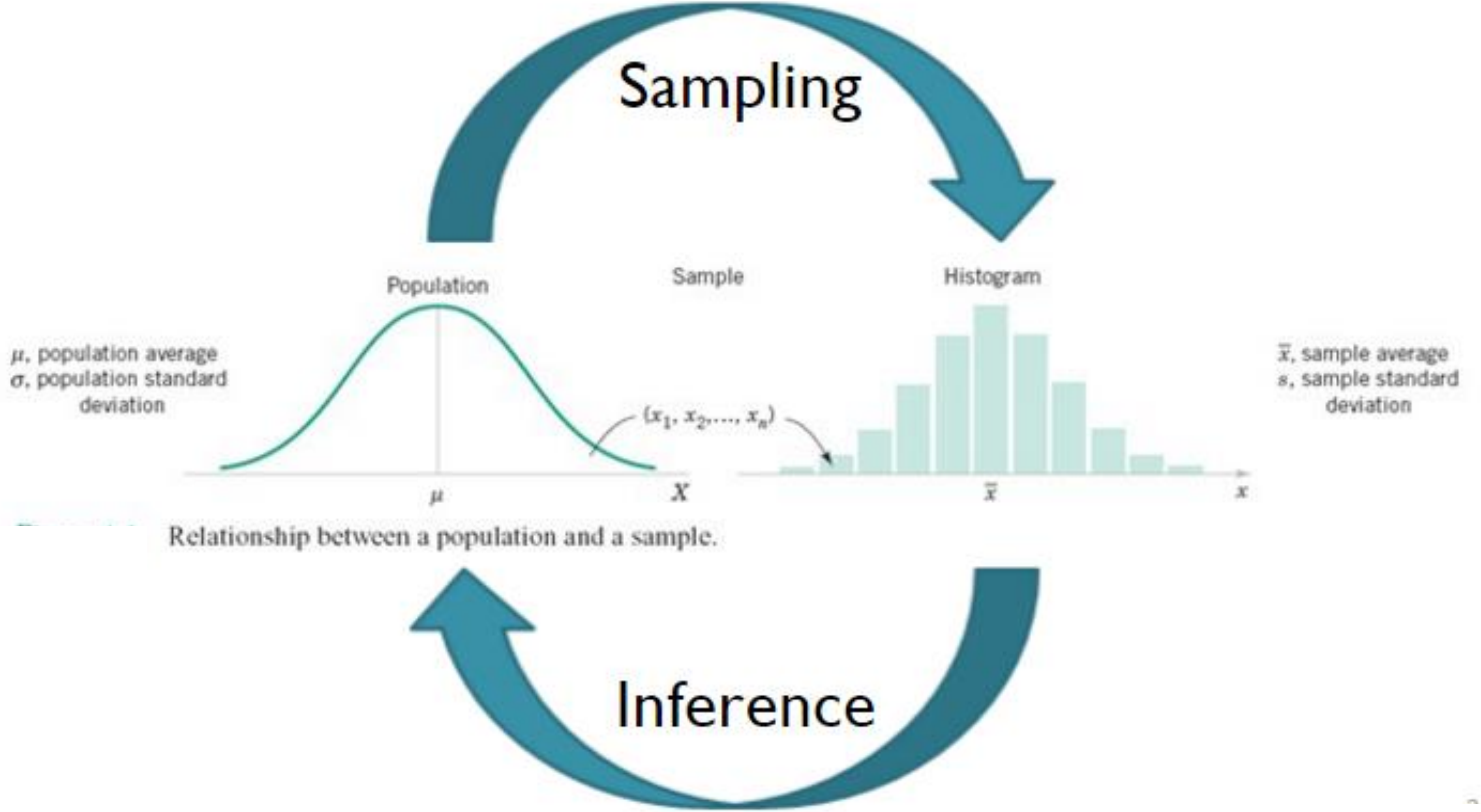
Dr. Dileepa Ediriweera

MBBS, MSc(Biostat), MSc(Biomedical Informatics), DCSD
FRSS(UK), MIAS(SL)

University of Kelaniya
Faculty of Medicine

Population vs Sample

- Population value vs Sample value
  - Parameter vs statistic
- Notations
  - Greek vs English



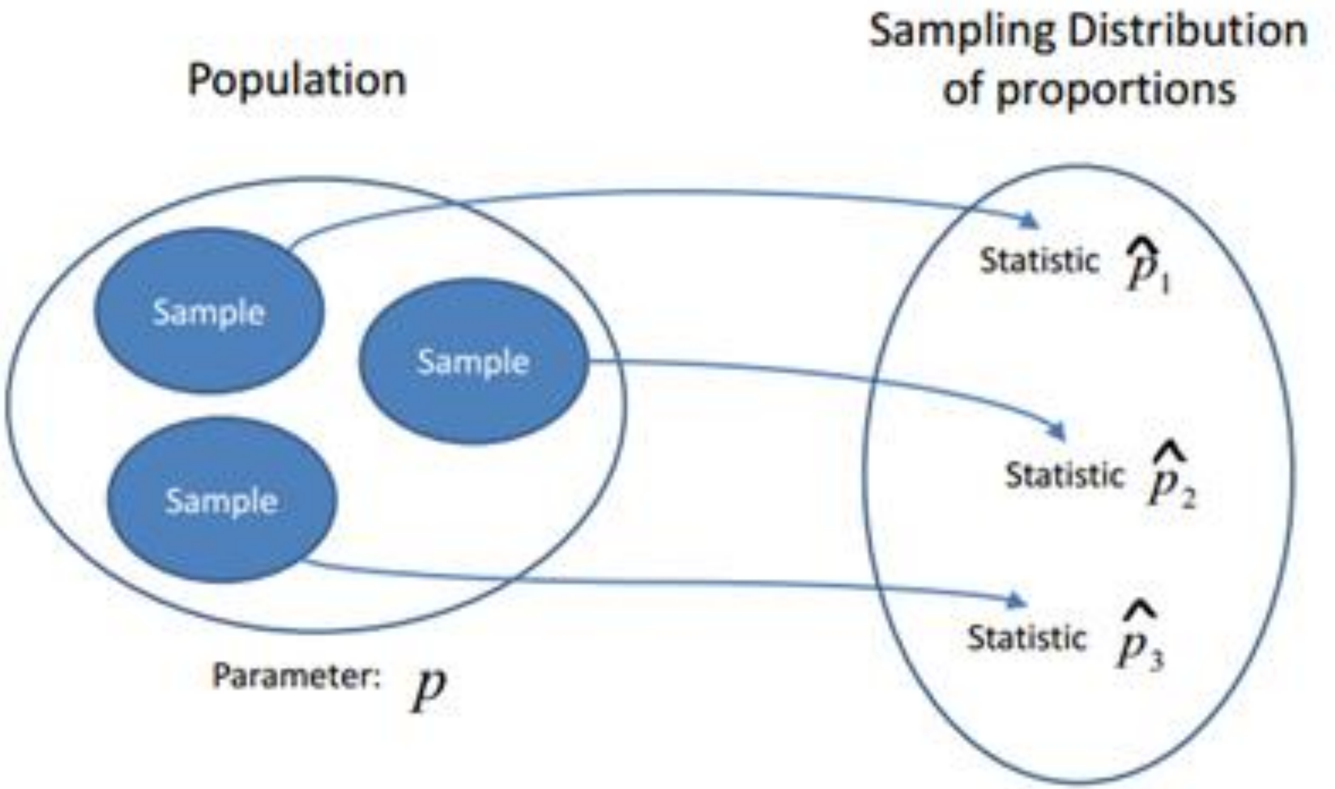
Sampling

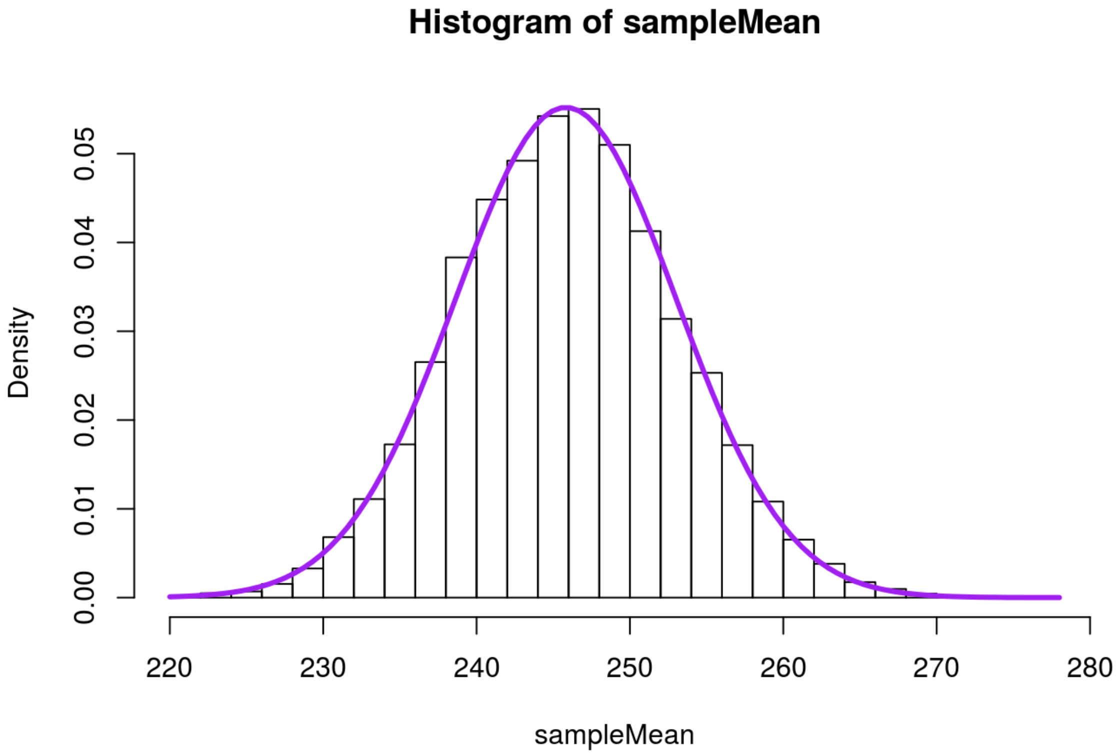
Inference

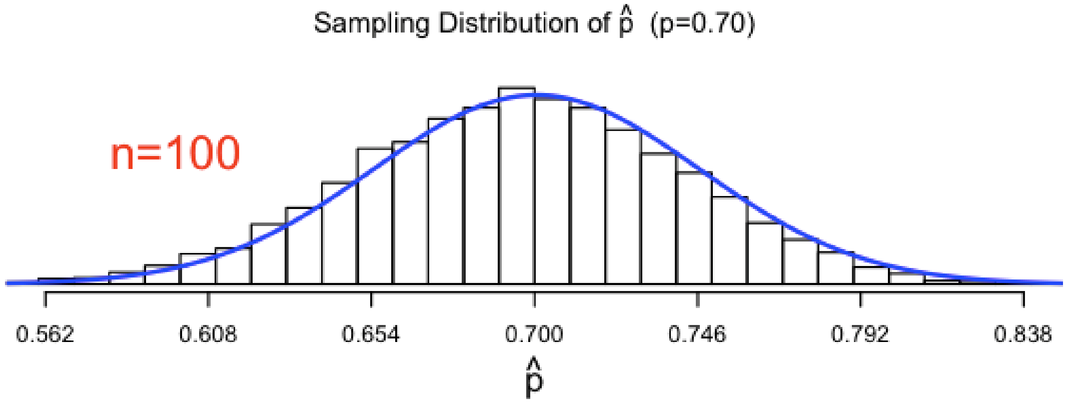
Relationship between a population and a sample.

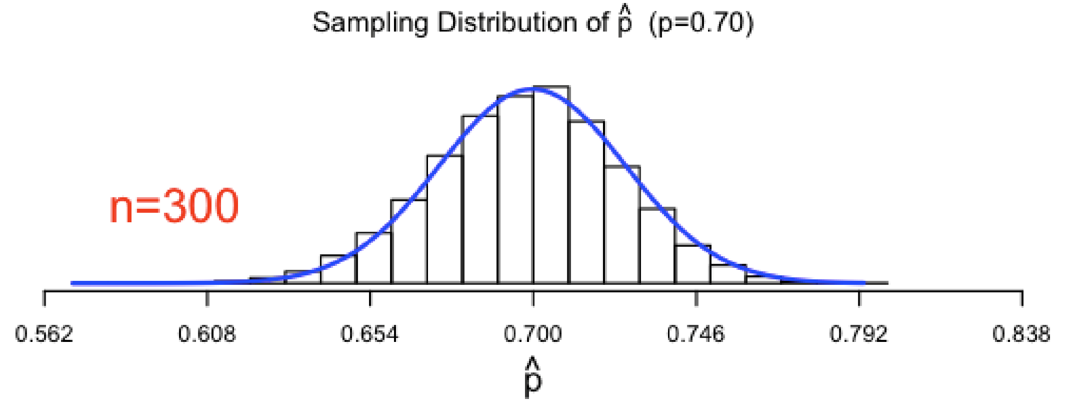
Population
Sample
Histogram
μ, population average
σ, population standard deviation
$(x_1, x_2, \dots, x_n)$
$\bar{x}$, sample average
s, sample standard deviation
μ
X
$\bar{x}$
x

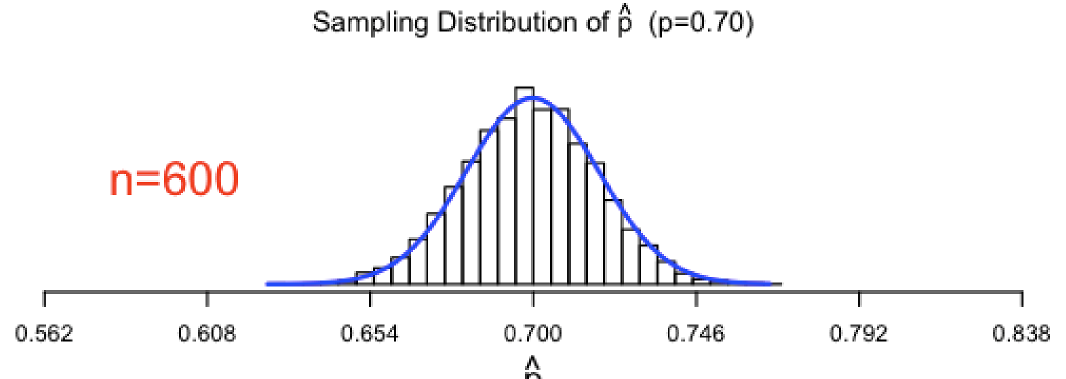
Sampling


Population
Sampling Distribution of proportions
Sample
Sample
Sample
Statistic $\hat{p}_1$
Statistic $\hat{p}_2$
Statistic $\hat{p}_3$
Parameter: p


Histogram of sampleMean
Density
0.00
0.01
0.02
0.03
0.04
0.05
220
230
240
250
260
270
280
sampleMean


Sampling Distribution of $\hat{p}$ (p=0.70)
n=100
0.562
0.608
0.654
0.700
0.746
0.792
0.838
$\hat{p}$


Sampling Distribution of $\hat{p}$ (p=0.70)
n=300
0.562
0.608
0.654
0.700
0.746
0.792
0.838
$\hat{p}$

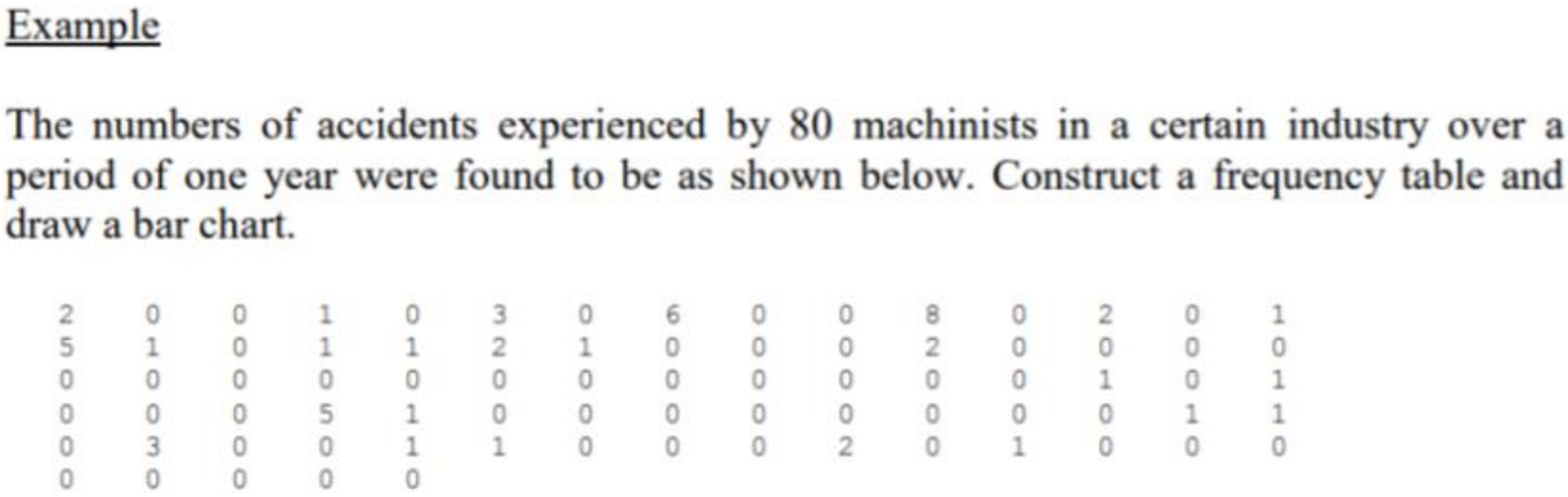

Sampling Distribution of $\hat{p}$ (p=0.70)
n=600
0.562
0.608
0.654
0.700
0.746
0.792
0.838
$\hat{p}$

Types of data

- Categorical (Qualitative, nominal)
  - Eg. Blood group
- Quantitative
  - Discrete – Eg. Number of students
  - Continuous – Eg. Height, Weight
- Ordinal (in-between case)
  - Eg. Exam grades (A, B, C, F)

Displaying data

- Discrete data: frequency table and bar chart



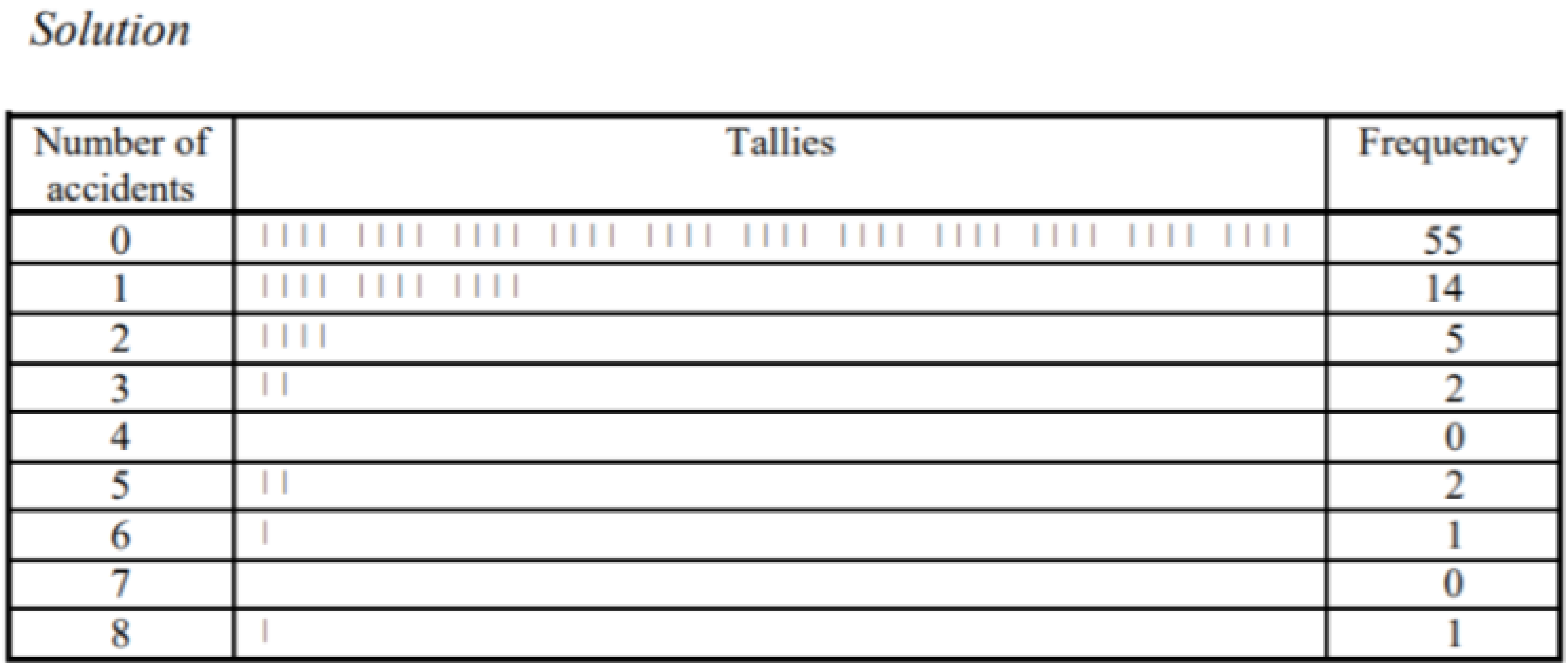
Example

The numbers of accidents experienced by 80 machinists in a certain industry over a period of one year were found to be as shown below. Construct a frequency table and draw a bar chart.

2	0	0	1	0	3	0	6	0	0	8	0	2	0	1
5	1	0	1	1	2	1	0	0	0	2	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
0	0	0	5	1	0	0	0	0	0	0	0	0	1	1
0	3	0	0	1	1	0	0	0	2	0	1	0	0	0
0	0	0	0	0										

Displaying data

- Discrete data: frequency table (සංඛ්‍යාත වගුව)



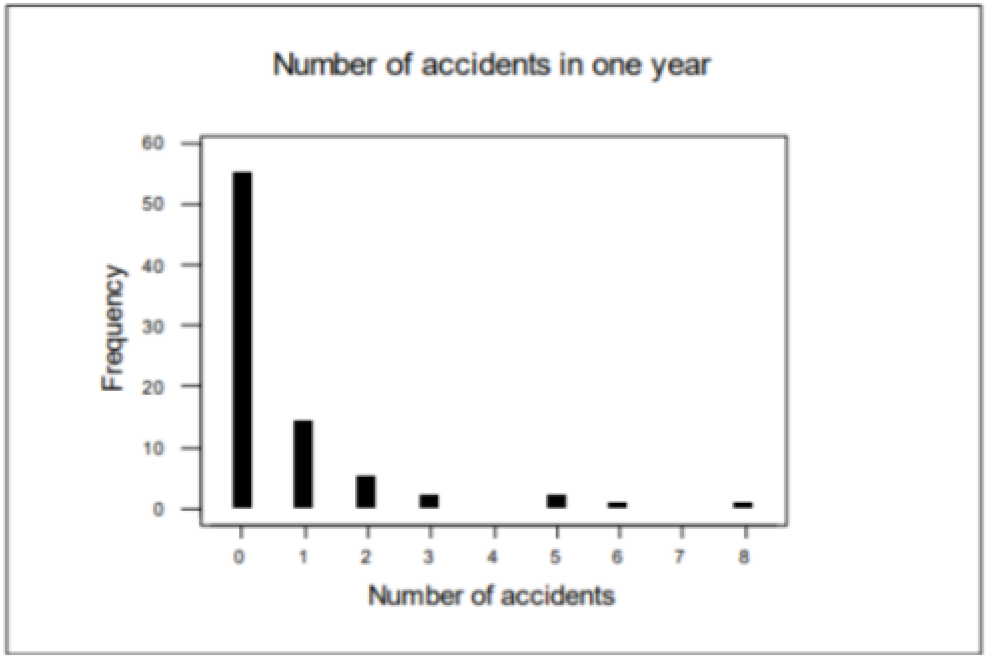
Solution

Number of accidents	Tallies	Frequency
0	|||| |||| |||| |||| |||| |||| |||| |||| |||| |||| ||||	55
1	|||| |||| ||||	14
2	||||	5
3	||	2
4		0
5	||	2
6	|	1
7		0
8	|	1

Displaying data

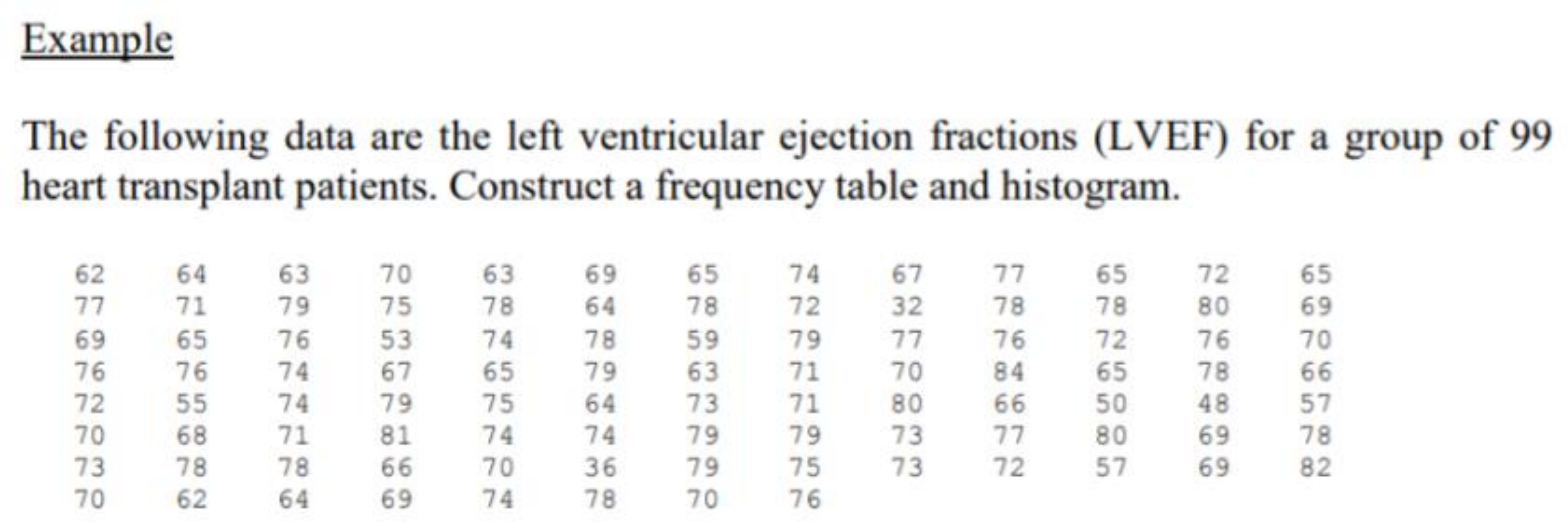
- Discrete data: bar char (තීරු සටහන්)

Barchart



Displaying data

- Continuous data: histograms (ජාල රේඛය)



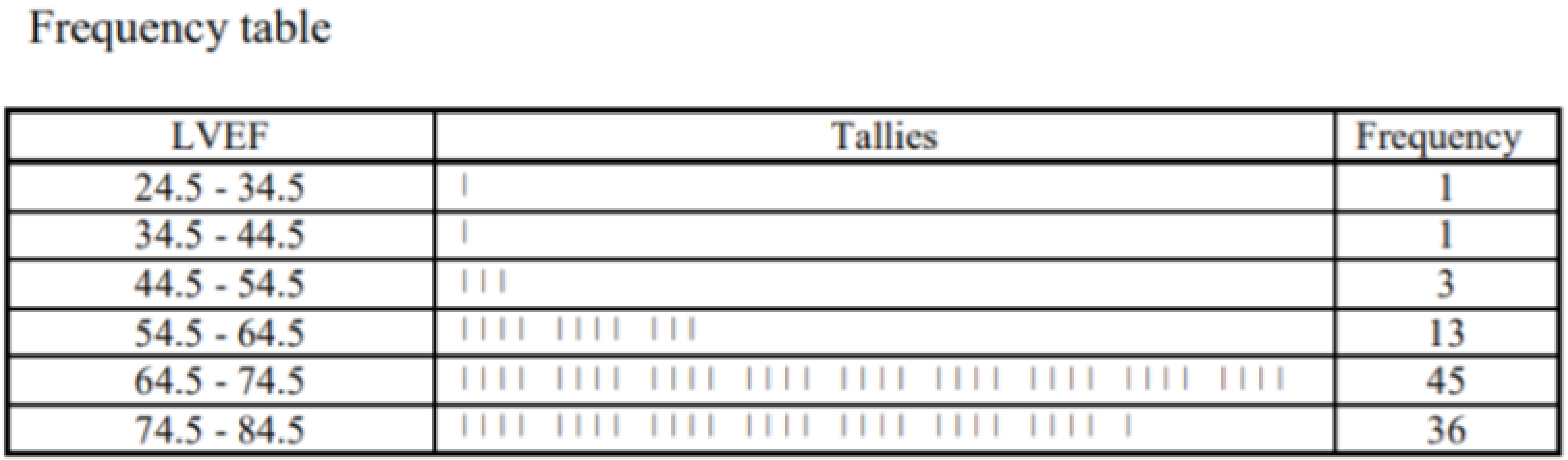
Example

The following data are the left ventricular ejection fractions (LVEF) for a group of 99 heart transplant patients. Construct a frequency table and histogram.

62	64	63	70	63	69	65	74	67	77	65	72	65
77	71	79	75	78	64	78	72	32	78	78	80	69
69	65	76	53	74	78	59	79	77	76	72	76	70
76	76	74	67	65	79	63	71	70	84	65	78	66
72	55	74	79	75	64	73	71	80	66	50	48	57
70	68	71	81	74	74	79	79	73	77	80	69	78
73	78	78	66	70	36	79	75	73	72	57	69	82
70	62	64	69	74	78	70	76					

Displaying data

- Continuous data: histograms



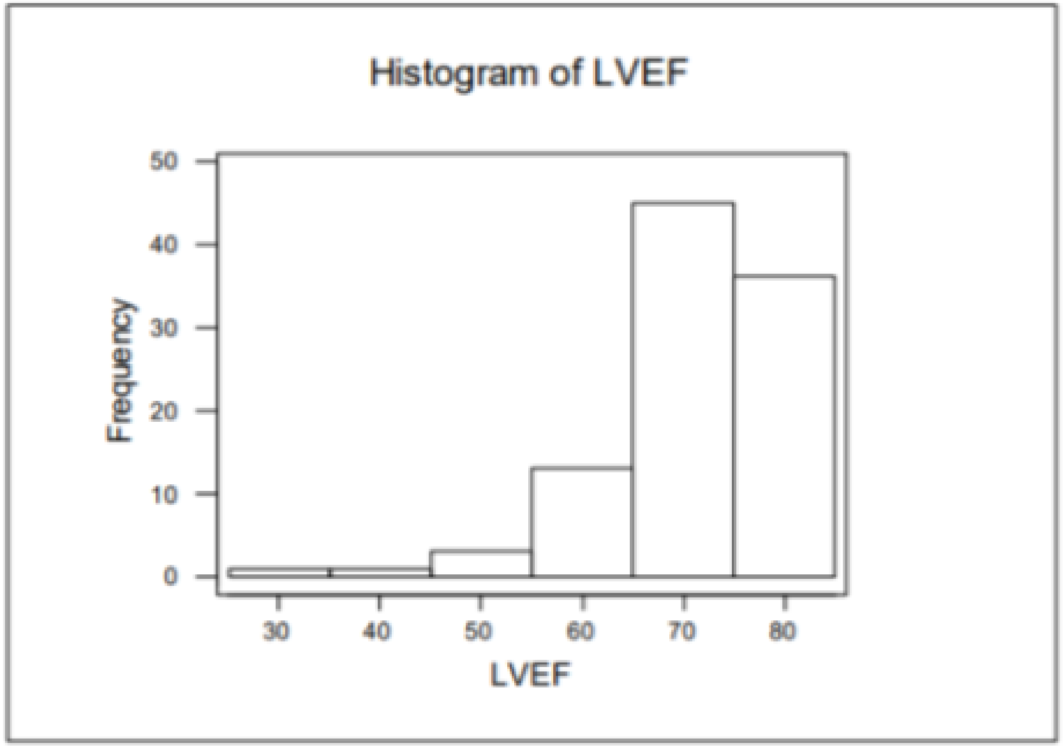
Frequency table

<table>
<tr><th>LVEF</th><th>Tallies</th><th>Frequency</th></tr>
<tr><td>24.5 - 34.5</td><td>|</td><td>1</td></tr>
<tr><td>34.5 - 44.5</td><td>|</td><td>1</td></tr>
<tr><td>44.5 - 54.5</td><td>|||</td><td>3</td></tr>
<tr><td>54.5 - 64.5</td><td>|||| |||| |||</td><td>13</td></tr>
<tr><td>64.5 - 74.5</td><td>|||| |||| |||| |||| |||| |||| |||| |||| ||||</td><td>45</td></tr>
<tr><td>74.5 - 84.5</td><td>|||| |||| |||| |||| |||| |||| |||| |</td><td>36</td></tr>
</table>

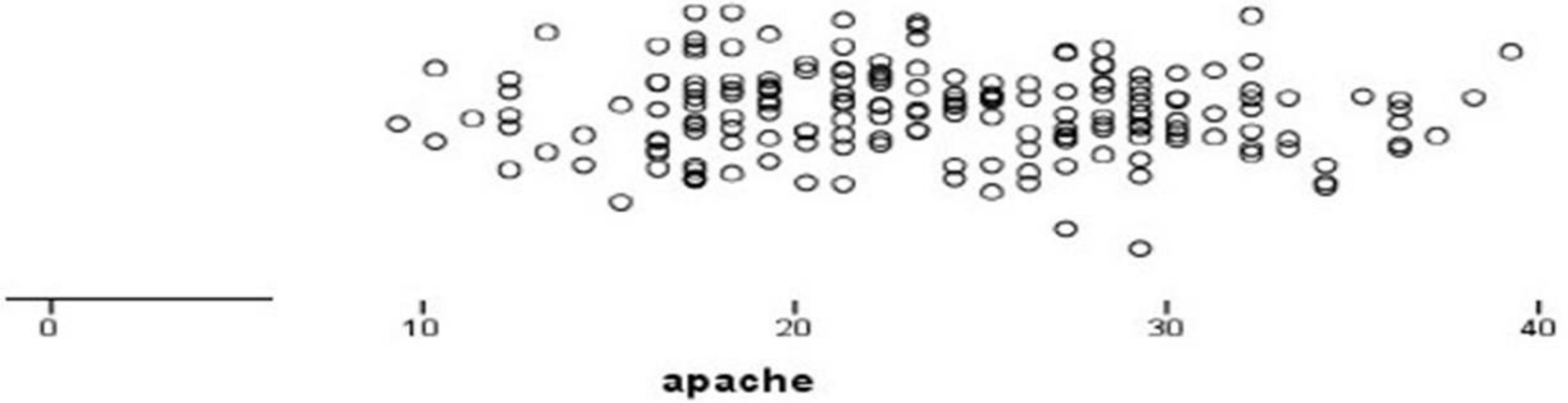
Displaying data

- Continuous data: histograms

Histogram



Summary Statistics



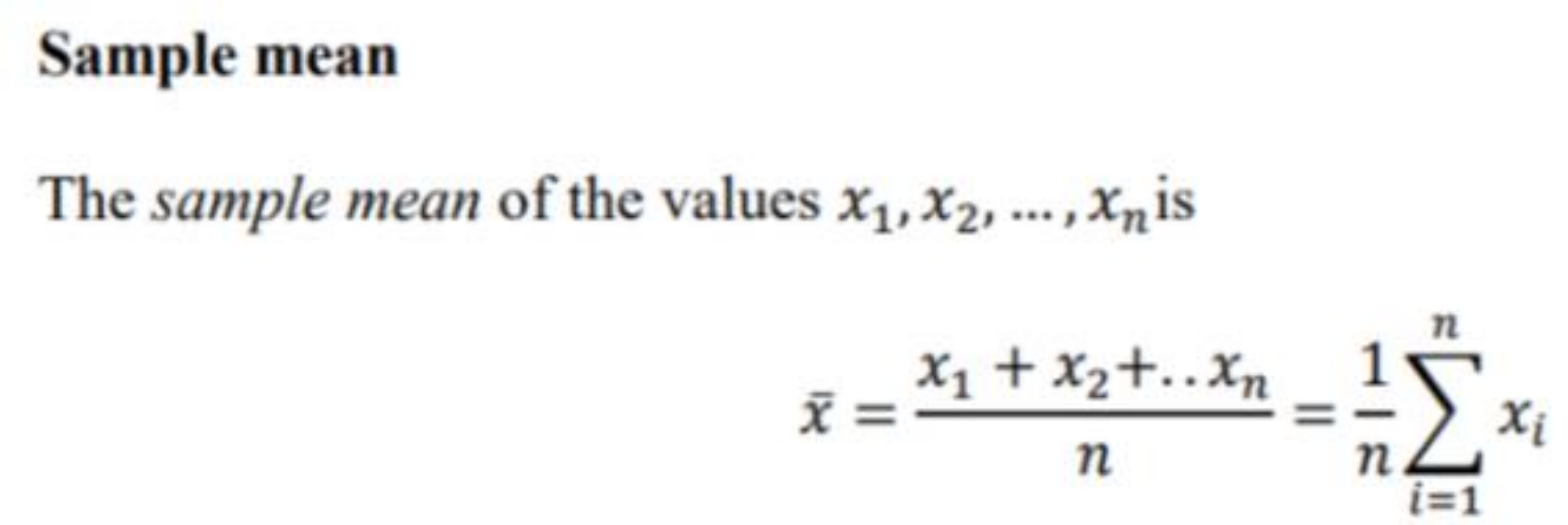
Central Tendency
Mean
Data Variation
Standard Deviation

Summary Statistics

- Measures of location (central tendency)
  - Sample mean (මධ්‍යනය)
  - Sample median (මධ්‍යස්ථය)
  - Mode (මාතය)

- Measure of dispersion
  - Range (පරාසය)
  - Standard deviation (සම්මත අපගමනය)

Sample mean



Sample mean

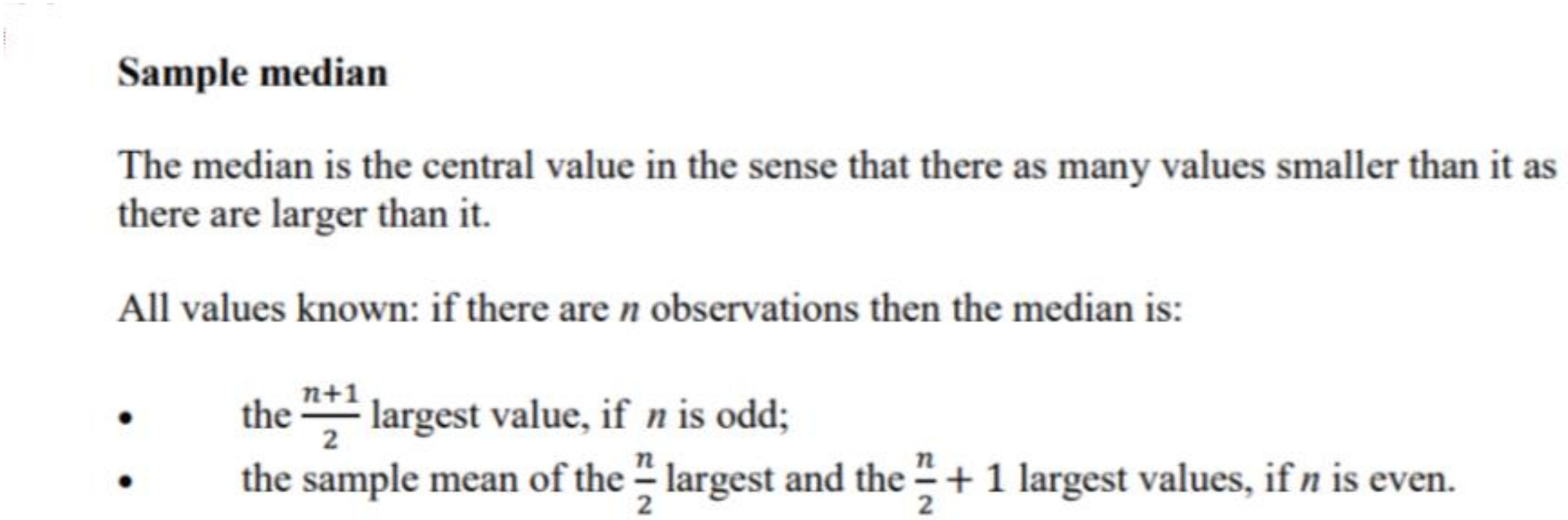
The *sample mean* of the values $x_1, x_2, \ldots, x_n$ is

$$\bar{x} = \frac{x_1 + x_2 + ..x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample mean

- How to calculate mean height of the class?

Sample median



Sample median

The median is the central value in the sense that there as many values smaller than it as there are larger than it.

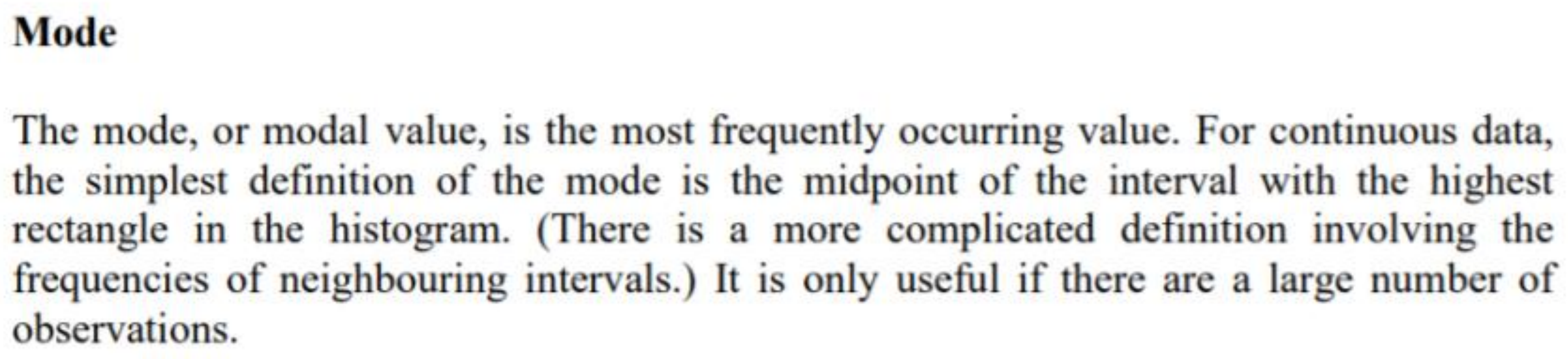
All values known: if there are n observations then the median is:

- the $\frac{n+1}{2}$ largest value, if n is odd;
- the sample mean of the $\frac{n}{2}$ largest and the $\frac{n}{2}+1$ largest values, if n is even.

Sample median

- How to calculate median height of the class?

Mode

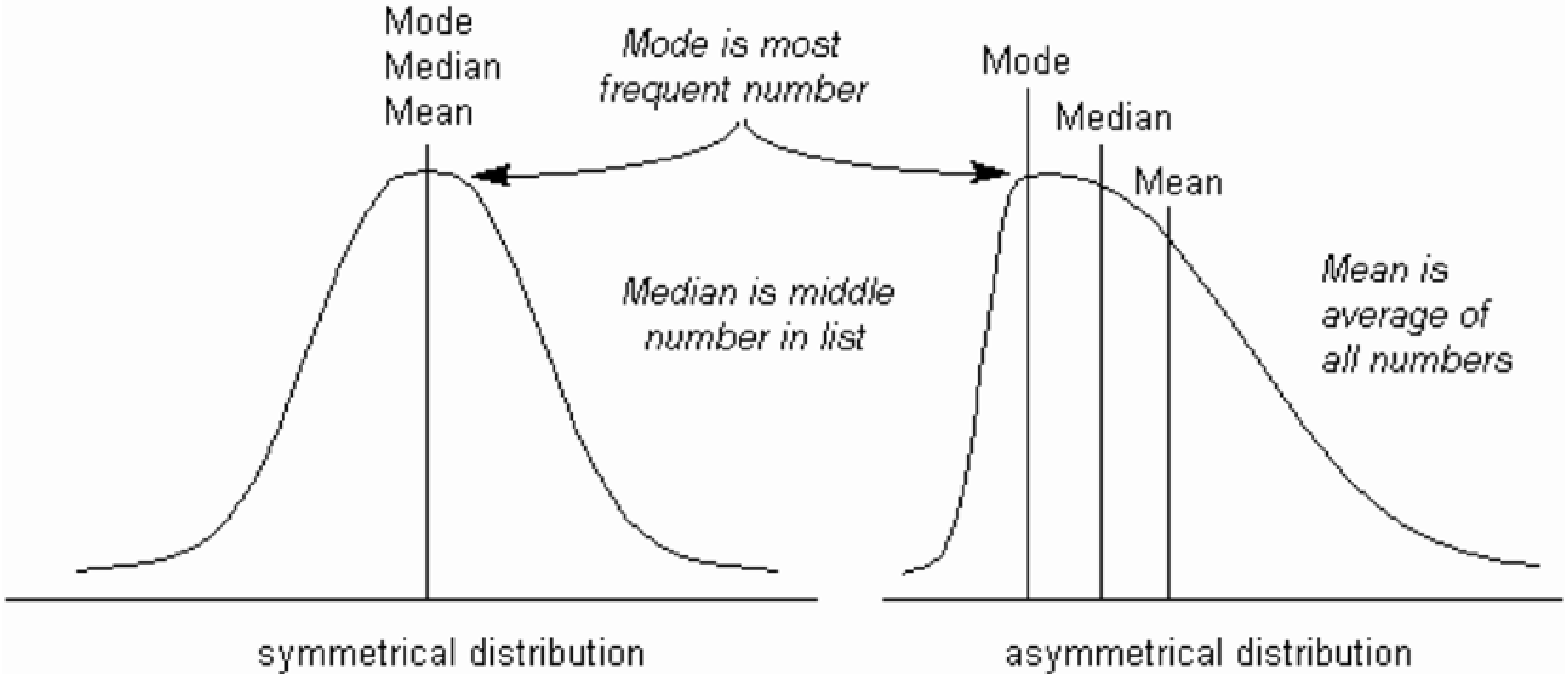


Mode

The mode, or modal value, is the most frequently occurring value. For continuous data, the simplest definition of the mode is the midpoint of the interval with the highest rectangle in the histogram. (There is a more complicated definition involving the frequencies of neighbouring intervals.) It is only useful if there are a large number of observations.

Mode

- How to find the mode of the height of this class?

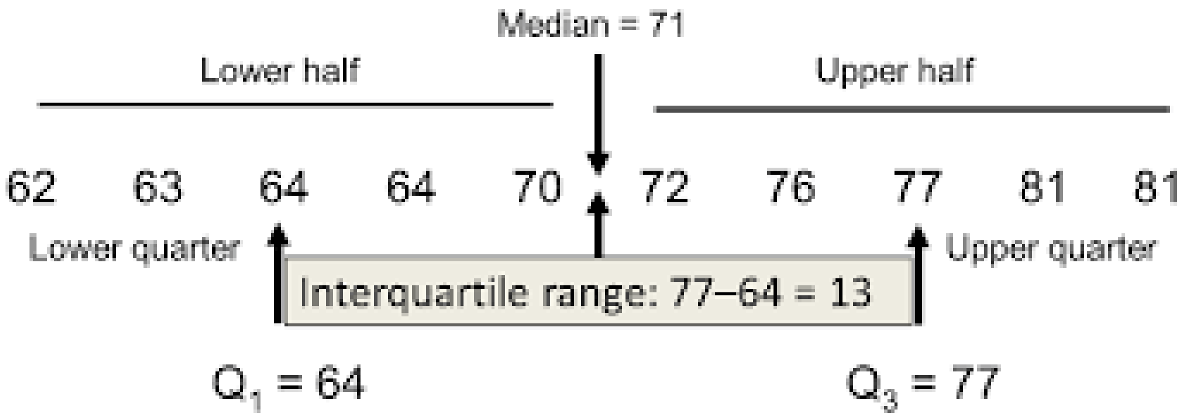

Mode
Median
Mean
Mode is most frequent number
Mode
Median
Mean
Median is middle number in list
Mean is average of all numbers
symmetrical distribution
asymmetrical distribution

Range and interquartile range

- Range
  - Difference between highest and lowest values

Range and interquartile range

- Interquartile range (අන්තශ්චතුර්ථක පරාසය)
  - Difference between Q1 and Q3
  - Q1: 25th percentile of the data. (splits off the lowest 25% of data from the highest 75%)
  - Q2: median of a data set is equal to the 50th percentile of the data (cuts data in half)
  - Q3: is equal to the 75th percentile of the data. (splits off the lowest 75% of data from highest 25%)

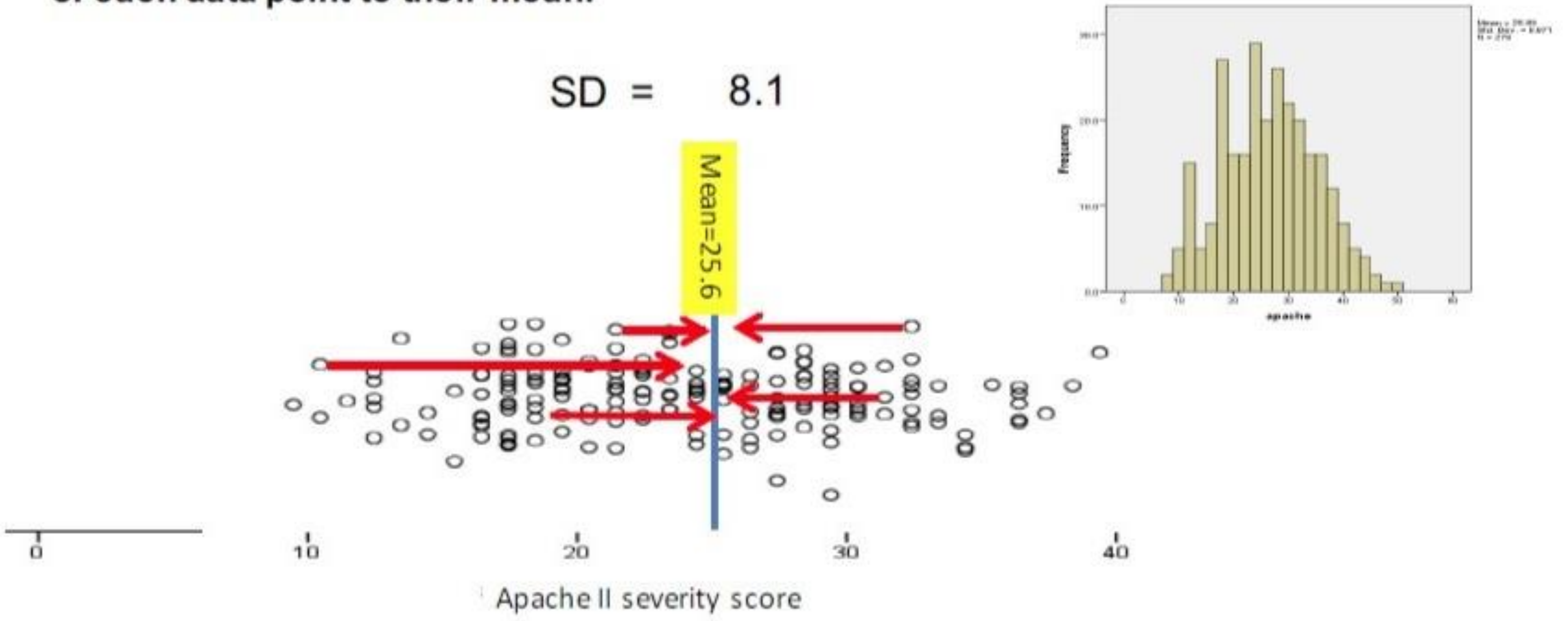


Ce

Standard deviation

Standard Deviation (SD)

SD describes "variation" of data, which is approximately equivalent with average distance of each data point to their mean.



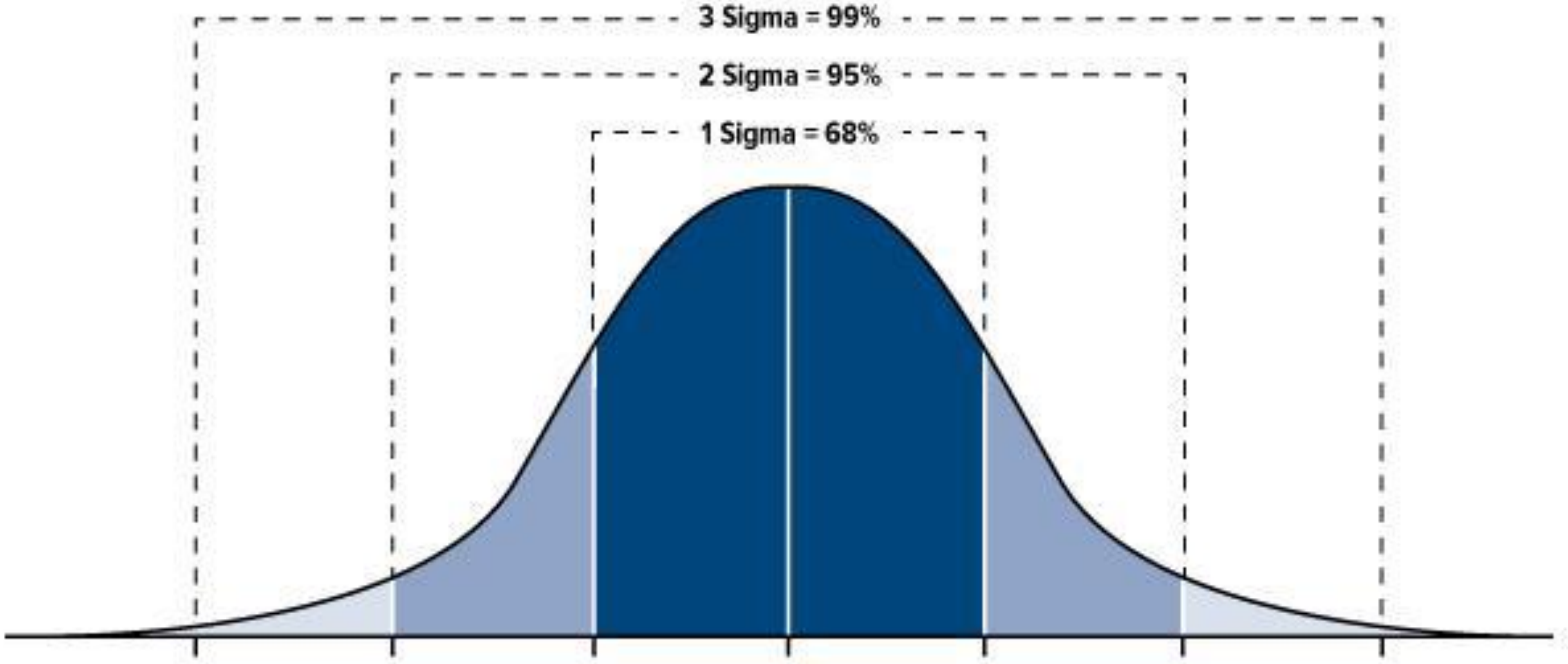
SD = 8.1

Standard deviation

$$SD = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

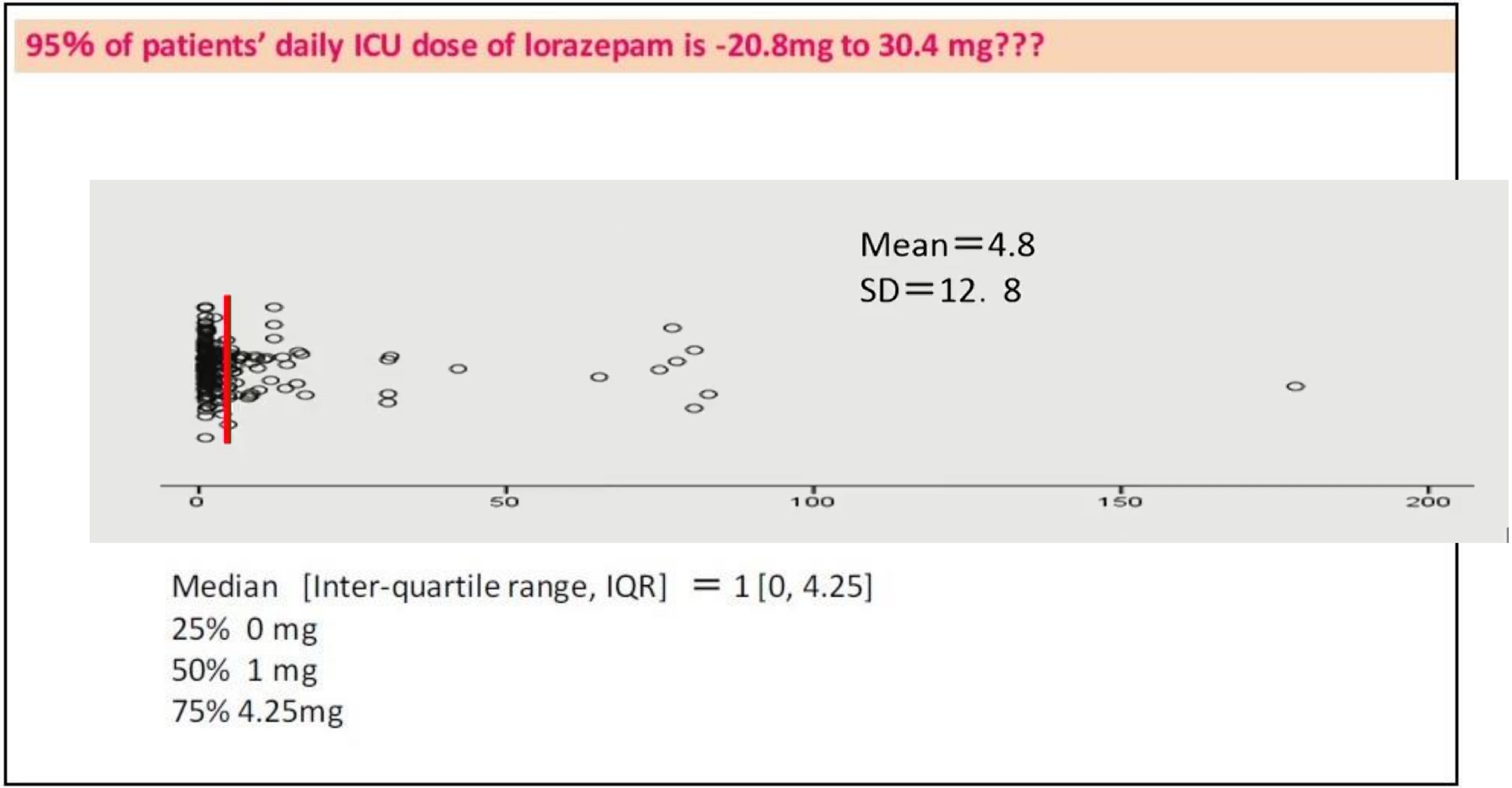
Why SD?

Standard Deviation (Sigma) Measures Degree of Variance from Average



Source: U.S. Global Investors

- Use of Inter-quartile range
- Mean and SD **not appropriate**



95% of patients' daily ICU dose of lorazepam is -20.8mg to 30.4 mg???

Median [Inter-quartile range, IQR] = 1 [0, 4.25]
25% 0 mg
50% 1 mg
75% 4.25mg

Feedback

1. what did you like about this session?
2. what didn't you like about this session?
3. what did you learn from this session?

Thank you